

**THE EFFECTS OF NO CHILD LEFT BEHIND ON STUDENT PERFORMANCE
IN ALABAMA'S RURAL SCHOOLS**

KINNUCAN, Henry W.*

SMITH, Martin D.

ZHENG, Yuqing

LLANES, Jose R.

Abstract

County level data for the period 1999-2007 are used to assess the effects of NCLB on student performance in Alabama's rural schools. Results suggest revisions to the state's accountability system associated with the Act had a positive effect on 8th grade test scores for language, and for test score gains in language between the 4th and 8th grades. However, the measured effects on test scores for reading and math are mostly zero or negative. This suggests NCLB failed in its major objective, which was to enhance students' proficiency in math and reading.

JEL classification: I2**Key words:** *dynamic panel data model, education production function, No Child Left Behind, rural schools, student performance*

1. Introduction

The No Child Left Behind Act (NCLB) of 2001 “[compels] states to conduct annual student assessments linked to state standards, to identify schools that are failing to make ‘adequate yearly progress’ (AYP) towards the stated goal of having all students achieve proficiency in reading and math by 2013-14 and to institute sanctions and rewards based on each school’s AYP status” (Dee and Jacob 2009, p. 2). The purpose of the present research is to determine the effectiveness of the Act by focusing on 4th and 8th grade test scores in Alabama's county schools. Alabama's county schools are a useful case study in that minorities and the economically disadvantaged -- a main target of the legislation -- are well represented. Nonwhite students constitute 39% of the schools' average enrollment, and the poverty rate across the 67 counties is 25% (Appendix 1 table A1A).

Thanks in part to a court order (now expired) that required the state to provide “adequate and equitable” funding to all school districts, several studies have been conducted to determine the likely impacts of increased per-pupil spending on student performance (Ferguson and Ladd 1996; Kinnucan, Zheng, and Brehmer 2006). Thus, a research base exists in which to inform model specification, and to compare results.

We focus on 4th and 8th grade test scores to be consistent with these earlier studies, but also because national assessments of NCLB by Lee (2006) and by Dee and Jacob (2009, 2011) are based on 4th and 8th grade scores. Lee (2006, p. 10) states “Previous studies are limited because they rely on states' own assessment results only “after” NCLB was adopted to show changes in achievement. Any change we see in student achievement after NCLB may reflect a continuing trend that occurred before NCLB.” In our data, 8th

* Kinnucan, email: kinnuhw@auburn.edu; and Smith are professor and former graduate research assistant, respectively, in the Department of Agricultural Economics and Rural Sociology, Auburn University, Auburn, Alabama, 36849-4201. Zheng is a research associate in the Dyson School of Applied Economics and Management, Cornell University, Ithaca, New York, and Llanes is a professor in the Department of Educational Foundations, Leadership and Technology, Auburn University. Appreciation is expressed to Patricia Duffy for helpful comments. Responsibility for final content, however, rests strictly with the authors.

grade test scores *declined* in the post-NCLB period (see Appendix 1, table A1B). Thus, the spurious trend problem identified by Lee is not an issue.

Finally, as noted by Dee and Jacob (2011, p. 419) “Although the reauthorization of NCLB is currently under consideration, the empirical evidence on the impact of NCLB on student achievement is, to date, extremely limited.” Of the studies that have been done (see, e.g., Dee and Jacob, 2011 and the references therein), none provides information on rural schools, the main focus of this research.¹

The analysis is based on county-level data for the period 1999-2007. A unique aspect of the analysis is that the education production function is specified to permit separate estimates of the short- and long-run impacts of NCLB. The dynamic specification permits test scores to adjust gradually to the accountability system, as contemplated by the AYP provision. The model is estimated using instrumental variables to correct for potential biases associated with endogenous school inputs.

In the next section we review the literature. The state’s accountability system is then discussed, followed by presentation of the model and empirical results. The paper concludes with a brief summary of the key findings.

2. Literature review

Student performance typically is evaluated using a production function, a cost function, or a supply-demand framework. Reviews of this literature are provided by Hanushek (1986, 1997, 2004), Levačić and Vignoles (2002), and Todd and Wolpin (2003). A review of the related literature on school accountability is provided by Figlio and Ladd (2008). In this review, we focus on the analyses by Ferguson and Ladd (1996) and Kinnucan, Zheng, and Brehmer (2006), the only known studies of student performance in Alabama’s public schools at the K-12 level.

In their analysis of test scores for all public schools (urban and rural), Ferguson and Ladd (hereafter “FL”) estimate an education production function using both student-level and district-level data. In their district-level model, FL specified 8th and 9th grade average test scores as a function of 3rd and 4th grade average scores. Hence, the estimated function is a growth or “value-added” model.² The model was estimated using data for 127 public school districts in Alabama for the 1989-90 school year. The data were weighted by the square-root of the school district’s enrollment to correct for heteroskedasticity.

FL found that the three most important variables explaining predicted test score gains were 3rd and 4th grade test scores in reading and mathematics (31 percent), the percentage of adults in the district with 16 or more years of schooling (17 percent), and teachers’ American College Testing (ACT) scores (15 percent). The remaining variables in the model, which included control variables to account for income, poverty, race,

¹ In Alabama the K-12 public school system is divided into 65 city school systems and 67 county school systems (Alabama Department of Education, 2007, p. 12). According to the U.S. Census Bureau, 72 percent of Alabama’s counties are rural (table 1). Hence, without too much imprecision, Alabama’s county schools may be considered rural schools.

² As Ferguson and Ladd (1996, p. 281) note, a true value-added model would require test scores for the same students between the two grade levels. Because these data were unavailable, Ferguson and Ladd used the 3rd and 4th grade test scores of current students as a proxy for what 8th and 9th graders in the same school would have earned when they were in the 3rd and 4th grades. For a good recent discussion of value-added models, see Hanushek and Rivkin (2010).

urbanization, and school district size, accounted for only 37 percent of the total explained variation. Of the four school inputs included in the model (teachers' ACT scores, percent of teachers with master's degrees, percent of teachers with five or more years of experience, and average class size), all but one (teachers' experience) were *statistically* significant. However, only ACT scores were *empirically* significant. In particular, class size, the most costly input to adjust owing to its lumpiness, accounted for only 7.5 percent of the explained variation.

To determine how spending might affect test scores, FL re-estimated the model with the four school input variables replaced by a single variable to indicate per-pupil spending on instruction. Importantly, the model was estimated using instrumental variables to account for the endogeneity of the spending variable, to wit (Ferguson and Ladd, 1996, p. 285):

In Alabama, there is compensatory spending in districts where student performance on standardized tests is the poorest. Thus, not only does spending affect student performance, but student performance affects spending. As a result, no systematic relationship emerges between test scores and actual instructional spending.

Using property value per student, per capita income, school district enrollment, and number of schools in the district as instruments, FL find a positive and significant relationship between predicted instructional spending and test scores. Moreover, the effect is non-linear, with the greatest benefit accruing to districts where per-pupil spending is the lowest, and no benefit to districts where instructional spending is above the median. These results, coupled with the findings from the model that included school inputs, led FL (1996, p. 288) to conclude that "money matters...especially for low-spending districts."

In their analysis of test scores in county schools (herein defined as rural schools), Kinnucan, Zheng, and Brehmer (hereafter "KZB") estimated a supply-demand model. The model consists of six equations describing the supply of education services, the demand for education services, the teacher-student ratio, teacher pay, federal funding, and market equilibrium. The quantity variable in the model is defined as average 8th grade test scores in language, reading and math. Two price variables are defined: a gross price that represents total per-pupil spending, and a net price that represents the difference total per-pupil spending and per-pupil state aid. State aid to the local school district, which accounts for 70 percent of total per-pupil spending, is treated as exogenous. The model was estimated using data for 67 county school districts for the 1992-93 school year. The endogenous variables in the model (test score, gross price, net price, federal funding, teacher-student ratio, and teacher pay) were instrumented using the exogenous supply and demand shifters in the model, augmented to include instruments used by FL. Following estimation, the model was solved for the reduced form and simulated to determine the net effect of changes in the seven exogenous variables in the model on equilibrium values of the six endogenous variables.

Based on estimated reduced-form elasticities, KZB found that the three most important variables affecting test score level are the county's poverty rate (-0.32 to -0.38), state aid to the local school district (0.20 to 0.29), and median family income (0.17 to 0.26).³ Of

³ Numbers in parentheses are reduced-form elasticities taken from table 5 of KZB's paper.

the remaining exogenous variables in the model, three had statistically significant effects on equilibrium test scores, but none were empirically significant. Specifically, the percent of adults in the county with a college degree had no effect on equilibrium test scores, while the percentage of students in county schools that are nonwhite (-0.004 to -0.012), the number of students enrolled in the county school system (-0.001 to -0.003), and the number of teachers with education beyond the Master's degree (0.034 to 0.040) had effects that were too small to matter.

Teacher pay, the most important shift variable in the supply equation, had an estimated structural elasticity of 3.02. This means that an isolated 1 percent increase in teacher pay would shift the supply curve for education services to the right from its initial equilibrium position by approximately 3 percent. By contrast, median family income, the most important shift variable in the demand equation, had an estimated structural elasticity of 0.67. Thus, the demand shift associated with a 1 percent change in median family income is only about 22 percent as large as the supply shift associated with a 1 change increase in teacher pay. Teacher-student ratio, a shift variable in the supply equation, had no effect on test scores. This result was not surprising, as the average class size in Alabama's county schools in 1992 was 17.5. This is well below the threshold of 23 identified by FL for learning gains to occur.

In KZB's model, increased state aid affects test scores indirectly through its effects on the equilibrium gross and net prices of educational services, but also through its effects on the equilibrium levels of teacher pay and class size (the inverse of teacher-student ratio). Of these two endogenous inputs, only teacher pay had a statistically significant effect on the supply of educational services. Consequently, KZB (2006, p. 487) conclude "increases in state aid can improve student performance *provided the incremental funding goes to teacher salaries and not to reductions in class size*" (emphasis added). Because state aid accounted for 70 percent of total per-pupil spending, KZB's findings are consistent with FL's conclusion that "money matters," with the caveat that added monies go to teacher pay and not to other inputs that may have little or no effect on student performance.

Combining the studies, teacher and family characteristics are important drivers of test scores in Alabama's county schools, and thus these variables need to be considered in any attempt to isolate the effects of NCLB. KZB's finding that poverty, family income, and teacher pay are germane to test scores in county schools is especially pertinent, as these variables moved in adverse directions during the NCLB period, which could explain the observed decline in test scores, a major hypothesis to be tested.

3. Alabama accountability

In 1995, the Alabama Legislature passed the *Education Accountability Plan* (Code of Alabama, 2001 Replacement). The plan, which went into effect in the 1995-96 school year, required all schools in the state be accountable in four areas: fiscal responsibility, school safety, discipline, and student achievement. More information in Appendix 2.

4. Model, estimation procedures and results

In this study, we follow FL and base our analysis on an education production function. Adopting Hanushek and Raymond's (2005) notation, the basic model in linear form is:

$$O_{ct} = \beta_0 + \beta_X X_{ct} + \beta_R R_{ct} + \gamma A_t + \rho_c + \varepsilon_{ct} \quad (1)$$

where O_{ct} is the level of student outcomes in county c in year t ; X_{ct} is a vector of family and nonschool inputs; R_{ct} is a vector of school resources; A_t is a measure of an accountability system (assumed to be the same across all counties) implemented to affect incentives and thus student performance; ρ_c is a fixed effect that captures policies and other unobserved effects at the county level; and ε_{ct} is an idiosyncratic error.

An implicit assumption underlying equation (1) is that student outcomes respond to changes in the accountability system and other explanatory variables in the same year in which the changes occur. The AYP provision in the NCLB Act contemplates it will take time for schools to adjust. Hence, the static specification given by equation (1) is inadequate, and an appropriate dynamic specification must be found.

For this purpose, we adopt the partial adjustment model (Griliches 1967; King and Thomas 2006). An advantage of the PAM over other possible dynamic specifications is that the short-run effects of policy shocks are easily distinguishable from their long-run effects.⁴ The first step in implementing the PAM is to rewrite equation (1) as:

$$O_{ct}^* = \beta_0 + \beta_X X_{ct} + \beta_R R_{ct} + \gamma A_t + \rho_c + \varepsilon_{ct} \quad (2)$$

where O_{ct}^* is the *desired* level of student outcomes in county c and year t .

Equation (2) is properly interpreted as the *long-run* education production function. The relationship between desired and actual student outcomes is given by the PAM as:

$$O_{ct} - O_{ct-1} = \lambda(O_{ct}^* - O_{ct-1}) \quad (3)$$

where λ is an “adjustment coefficient,” which is restricted to lie on the semi-closed interval $(0, 1]$. In essence, the PAM implies that yearly progress is proportional to the achievement gap, where the achievement gap is the difference between desired student outcomes in the current year, and actual outcomes in the previous year. The larger the achievement gap, the larger the adjustment in any given year. Alternatively, for any given achievement gap, the larger the adjustment coefficient, the greater the adjustment in the current year. If $\lambda = 1$, equation (3) reduces to $O_t = O_t^*$ and student outcomes adjust fully in the same year in which the policy or other shock occurs (static model).

The PAM may be written alternatively as:

$$O_{ct} = \lambda O_{ct}^* + (1 - \lambda) O_{ct-1} \quad (4)$$

This form shows that observed student outcomes in any given year lie between the desired outcomes for that year, and the actual outcomes in the preceding year. Substituting equation (2) into equation (4) yields:

$$O_{ct} = \lambda \beta_0 + \lambda \beta_X X_{ct} + \lambda \beta_R R_{ct} + \lambda \gamma A_t + (1 - \lambda) O_{ct-1} + \lambda (\rho_c + \varepsilon_{ct}). \quad (5)$$

Equation (5) is the *short-run* education production function. In particular, the accountability system’s short-run effect on student outcomes is given by $\lambda \gamma$. This coefficient tells the effect after one year. The corresponding long-run effect is given by γ , which is obtained by dividing the estimated coefficient of A_t by one minus the estimated coefficient of the lagged dependent variable (LDV). A similar interpretation applies to the other variables in the model. Thus, for example, the short-run effect of family and nonschool inputs on student outcomes is given by $\lambda \beta_X$ and the corresponding long-run effect is given by β_X . The latter effect is obtained by dividing the estimated

⁴ For a good recent discussion of these models, see Rothstein (2010).

coefficient of X_{ct} by one minus the estimated coefficient of the LDV. Because the adjustment coefficient is bounded between zero and one (inclusive), the short-run effect is never larger than the long-run effect. If the estimated coefficient of the LDV is zero, then $\lambda = 1$ and equation (5) reduces to a static model. That is, equation (5) nests equation (1).

Empirical specification and estimation procedures: The empirical model focuses on student performance in Alabama's 67 county school systems for the period 1999-2007. Fourth and 8th grade test scores from the Stanford Achievement Test (SAT) serve as the performance indicator. These scores are averages taken over all public schools in each county *exclusive of city schools*. The continuous variables are specified in log form to permit coefficients to be interpreted as elasticities. In the logarithmic specification λ is interpreted as the "elasticity of adjustment" rather than the "coefficient of adjustment" (Nerlove and Bessler 2001). The revised accountability system, i.e., the one that incorporates NCLB provisions into the state's original system, is modeled using a dummy variable that assumes the value of zero in the pre-NCLB period (1999-2002), and the value of one in the post-NCLB period (2003-2007).⁵ The NCLB dummy is interacted with dummy variables for the eight State Board of Education (SBE) school districts in Alabama to test whether NCLB effects differ across the SBE districts. (Actually, there are seven interaction terms, as District 5 is treated as the omitted category.) A map showing the SBE districts can be found on the Alabama Department of Education's web site. The X and R variables in the model match those specified by KZB in their analysis based on 1992 data. Hence, the present analysis can be viewed as an update of KZB's study to include new education policies at the federal level, but also dynamics. Two models are estimated:

Model A: test score level

$$\begin{aligned} \ln TS8_{ct} = & \pi_0 + \pi_{LDV} \ln TS8_{ct-1} + \pi_{NCLB_5} NCLB_t + \sum_{j=1, j \neq 5}^8 \pi_{NCLB_j} Dj_t \cdot NCLB_t \\ & + \pi_{INC} \ln INC_{ct} + \pi_{POV} \ln POV_{ct} + \pi_{RURAL} RURAL_{ct} + \pi_{RACE} \ln RACE_{ct} \\ & + \pi_{DSIZE} \ln ENROL_{ct} + \pi_{PPS} \ln PPS_{ct} + \pi_{TPAY} \ln TPAY_{ct} + \lambda(\rho_c + \varepsilon_{ct}) \end{aligned} \quad (6)$$

Model B: test score ratio

$$\begin{aligned} \ln(TS8_{ct}/TS4_{ct}) = & \omega_0 + \omega_{LDV} (TS8_{ct-1}/TS4_{ct-1}) + \omega_{NCLB_5} NCLB_t + \sum_{j=1, j \neq 5}^8 \omega_{NCLB_j} Dj_t \cdot NCLB_t \\ & + \omega_{INC} \ln INC_{ct} + \omega_{POV} \ln POV_{ct} + \omega_{RURAL} RURAL_{ct} + \omega_{RACE} \ln RACE_{ct} \\ & + \omega_{DSIZE} \ln ENROL_{ct} + \omega_{PPS} \ln PPS_{ct} + \omega_{TPAY} \ln TPAY_{ct} + \lambda'(\rho'_c + \varepsilon'_{ct}) \end{aligned} \quad (7)$$

⁵ Although the NCLB Act was signed into law by President George W. Bush on 8 January 2002, we define the NCLB era to begin in 2003. The reason is that a switching regression analysis of the data used in this study indicated the Act did not begin to take hold in Alabama until the 2003 school year at the earliest (Smith 2009), a result consistent with the fact that rules for implementing NCLB were not adopted by the state until June 2003. In preliminary analysis we experimented with alternative specifications of the NCLB variable, including a trend variable that assumed a value of zero prior to NCLB. The best results in terms of the models' explanatory power were obtained from the simple binary variable specification as defined in equations (6) and (7).

Definitions of the variables specified in each model along with their summary statistics are provided in table 1. A data appendix with sources is available upon request. The X variables that define family and nonschool inputs are *INC*, *POV*, *RURAL*, and *RACE*; the R variables that define school resources are *ENROL*, *PPS*, and *TPAY*. Class size is not included in the model because the average class size in Alabama's county schools over our sample period is 15 (standard deviation = 1.4). As mentioned earlier, this is well below the threshold of 23 identified by FL for learning gains to occur, a result confirmed in KZB's analysis. Also, *UNEMP*, a variable that was included in KZB's analysis as a demand shifter, is not included in the present model because KZB found it to be insignificant, a result confirmed in preliminary testing of the present model.

Model B differs from Model A in that the dependent variable is defined as the log ratio of 8th to 4th grade test scores. The motivation for Model B is that 8th grade test scores are lower than 4th grade test scores (see table 1), which suggests Alabama students lose ground relative to their peers nationwide as they move through the county school system. Model B, which was not estimated in KZB's study, provides a basis for assessing the determinants of this erosion, and whether NCLB reduced it.⁶

The models are estimated using test score data for reading, math, and language (six regressions altogether). Based on Hausman tests that indicated *ENROL*, *PPS*, and *TPAY* are endogeneous, the models were estimated using the Generalized Methods of Moments (GMM) estimator for Dynamic Panel Data models in EViews 7.0 (2009). The GMM estimator requires that the number of instruments be at least as large as the number of parameters to be estimated. The identified right-hand-sided endogenous variables were instrumented using i) lagged values for *ENROL*, *PPS*, *TPAY*, and the dependent variable, ii) current values for each of the remaining explanatory variables in the model, and iii) current values for *ED*, *PV*, and *UNEMP* (see table 1 for definitions). This yielded 19 instruments, two more than the number of parameters to be estimated. We comment preliminary analysis in Appendix 4.

Estimation results: As a first step in the estimation process, we tested three hypotheses regarding model specification:

⁶ As alluded to earlier in connection with FL's study, Model B under certain conditions can be interpreted as a value-added or "growth" model. To see this, suppress the error term and write the dynamic production function in value-added form as

$\ln TS8_{ct} = a + b \ln TS4_{ct}^A + c \ln X_{ct} + d \ln R_{ct} + eA_t + f \ln TS8_{ct-1}$ where $TS4_{ct}^A$ is the actual test scores that current 8th grade students received when they were in the 4th grade. Subtracting $\ln TS4_{ct}^A$ from both sides of this equation yields

$\ln(TS8_{ct}/TS4_{ct}^A) = a + (b-1) \ln TS4_{ct}^A + c \ln X_{ct} + d \ln R_{ct} + eA_t + f \ln TS8_{ct-1}$. This equation reduces to Model B when $b = 1$ and $TS4_{ct}^A = TS4_{ct}$ where $TS4_{ct}$ are the observed test scores of current 4th graders. The $TS4_{ct}^A = TS4_{ct}$ restriction implies that test scores of current 4th graders are good proxies for the scores that current 8th graders received when they were in the 4th grade, a maintained hypothesis in FL's study. The restriction $b = 1$ converts the value-added model to a linear growth model. As noted by Meyer (1997, p. 298, fn. 6), an advantage of the linear growth model (as opposed to the more general model that retains $TS4_{ct}^A$ as a right-hand-side variable) is that estimation problems associated with measurement error in the prior test score are avoided.

Hypothesis 1: The coefficients of the lagged dependent variable are zero. This hypothesis asserts the static specification is adequate.

Hypothesis 2: The coefficients of the interaction terms involving NCLB are jointly zero. This hypothesis asserts the NCLB effect is uniform across the SBE districts.

Hypothesis 3: The coefficients of the NCLB variables are jointly zero. This hypothesis asserts NCLB had no effect on test scores.

Hypothesis 1 entails one restriction and thus is tested using a simple t -test. Hypotheses 2 and 3 entail 7 and 8 restrictions, respectively, and thus are tested using a Wald test. For the purpose of these tests, Model A is relabeled L1, L2, and L3 to correspond to equation (6) for reading, math and language, respectively, and Model B is relabeled R1, R2, and R3 to correspond to equation (7) for the same subjects.

Results, shown in table 2, indicate Hypothesis 3 can be rejected in all models with almost no probability of a Type I error ($p < 0.004$).

Model ^a	H _N : Static Model		H _N : No Interaction Effects		H _N : No NCLB Effects	
	t -ratio	p -value	Computed Chi-square	p -value	Computed Chi-square	p -value
L1	-0.63	0.5287	20.38 (7) ^b	0.0048	49.39 (8)	0.0000
L2	4.88	0.0000	19.92 (7)	0.0058	28.44 (8)	0.0004
L3	2.16	0.0317	13.63 (7)	0.0581	33.77 (8)	0.0000
R1	1.68	0.0941	22.11 (7)	0.0024	81.22 (8)	0.0000
R2	5.59	0.0000	22.83 (7)	0.0018	22.83 (8)	0.0036
R3	2.31	0.0215	15.11 (7)	0.0347	70.74 (8)	0.0000

^a Models L1, L2, and L3 refer to 8th grade test-score *level* equations for reading, math, and language, respectively. Models R1, R2, and R3 refer to 8th/4th grade test-score *ratio* equations for reading, math, and language, respectively. For full regression results, see appendix tables 2 and 3.^b Number in parenthesis is degrees of freedom (= number of zero restrictions).

Thus, the NCLB variables belong in the specification. Hypothesis 2 is rejected at the $p < 0.05$ level in all models except L3, where rejection occurs at the $p < 0.06$ level. Hence, NCLB effects appear to differ across the SBE districts. Hypothesis 1 is rejected at the $p < 0.10$ level in all models except L1, and is rejected at the $p < 0.05$ level in all models except L1 and R1. Thus, the static specification is rejected at usual probability levels. Combining these results, equations (6) and (7) are compatible with the data in the sense that the considered restricted forms would have lead to specification error.

Estimation results are satisfactory in that the models show good explanatory power and most of the estimated coefficients for the X and R variables are significant and have signs that are consistent with KZB. (See appendix tables A2 and A3 for details.) The adjusted R^2 s for Model A range from 0.83 for language to 0.90 for reading; the corresponding range for Model B is 0.61 for language to 0.66 for math. The J -statistics for the models are less than 4.0, which, by the Sargan test, imply that the over-identifying restrictions are valid. Importantly, the point estimates for the LDV across all models have an upper bound of 0.24, which suggests adjustments to policy and other shocks are completed in

three years or less.⁷ A “persistence parameter” of 0.24 also implies that long-run elasticities are 24 percent larger than short-run elasticities, or, equivalently, that 76 percent of the total response is completed in the first year. This suggests interventions to improve student performance in Alabama’s county schools can produce results relatively quickly, in three years or less. The remaining discussion focuses on elasticities, as these are germane to policy.

Elasticity Estimates. The elasticities associated with the dummy variables in the model (the *NCLB* terms and *RURAL*) are semi-elasticities computed using the procedure suggested by Halvorsen and Palmquist (1980).⁸ Focusing first on the accountability system, the effects of NCLB are modest, confined to selected SBE districts, and not always positive (table A6 in Appendix 3).

Simulation The analysis thus far suggests *DSIZE*, *INC*, *POV*, and *TPAY* are the most important variables affecting student performance in Alabama’s county schools. However, as noted by Duffy (1989) in another context, the empirical importance of a variable depends not only on the size of its elasticity, but also on the magnitude of the movement in the corresponding variable. To gain a clearer picture of the relative importance of these variables, we “simulated” the estimated models by multiplying their respective long-run elasticities by the percentage changes in each of the variables between the pre- and post-NCLB periods as shown in tables A4 and A5 in Appendix 1.

NCLB Recap: The estimated effects of NCLB across the eight State Board of Education districts are summarized in table 6.

Nonwhite enrollment percentages, poverty rates, and enrollment shares for each district are included in the table to provide context. The NCLB effects in the table are taken from table 3, with statistically insignificant effects set to zero. The district-specific effects are multiplied by enrollment shares and summed get a statewide effect. Based on this procedure, NCLB is associated with statewide declines in average test score levels for reading and math of 2.6 percent and 0.6 percent, respectively, and a statewide increase in test score level for language of 3.2 percent. For test score ratio, NCLB is associated with a statewide decline in reading of 4.7 percent, and statewide increases in math and language of 0.1 percent and 8.1 percent, respectively. Thus, at the state level NCLB is associated with an improvement in test scores for language, but at the expense of test scores for reading.

⁷Letting $\hat{\delta}$ equal the estimated coefficient of the lagged dependent variable, the adjustment interval may be computed using the formula $\hat{\delta}^N = (1 - A)$ where N is the number of periods required to achieve $A/100$ percent adjustment to the new equilibrium. Setting $\hat{\delta} = 0.24$ and solving for N with $A = 95\%$ yields $N = \frac{\ln(0.05)}{\ln \hat{\delta}} = \frac{\ln(0.05)}{\ln(0.24)} = 2.1$. This suggests test scores achieve 95 percent of their full adjustment to a policy shock in a little over two years. 99 percent adjustment takes 3.2 years.

⁸Letting $\hat{\theta}$ equal the estimated coefficient of the dummy variable in a log-linear model, the semi-elasticity is computed using the formula $E = \exp(\hat{\theta}) - 1$. The semi-elasticities in table 3 were obtained by dividing the estimated thetas by one minus the estimated coefficient of the LDV prior to insertion in the formula. Hence, the semi-elasticities are long run.

Table 6. Nonwhite enrollment, poverty rate, enrollment share, and estimated effects of NCLB on test scores by State Board of Education (SBE) districts, Alabama county schools, 1999-2007

SBE District	Percent Nonwhite Enrollment ^a	Poverty Rate ^a	Enrollment Share ^a	8 th Grade Test Level			8 th /4 th Grade Test Score Ratio		
				Reading	Math	Language	Reading	Math	Language
1	39.5	22.8	0.193	0	0	0.059	-0.040	0	0.111
2	33.1	23.6	0.095	-0.029	0	0	0	0.052	0.089
3	21.8	19.7	0.152	0	0	0.054	-0.047	0	0.074
4	44.4	24.1	0.093	-0.068	-0.066	0	-0.041	0	0.123
5	74.8	31.7	0.187	-0.064	0	0	-0.124	-0.069	0.036
6	5.9	19.4	0.064	-0.070	0	0	-0.077	0	0
7	14.1	22.6	0.116	0	0	0.071	0	0.079	0.103
8	15.8	19.2	0.100	0	0	0.044	0	0	0.093
All	39.1	24.8	1.000	-0.026 ^b	-0.006 ^b	0.032 ^b	-0.047 ^b	0.001 ^b	0.081 ^b

The statewide effects of NCLB on math scores is basically nil. A major goal of NCLB was to level the academic playing field for minorities and the economically disadvantaged. There is little in the SAT data to suggest much leveling has occurred with respect to Alabama's county schools. District 5, the blackest district in the state with 74.8 percent nonwhite enrollment and the poorest district with a poverty rate of 31.7 percent, shows a positive effect of NCLB on test score gain for language. But the positive effect for language is erased by negative effects in reading and math. In fact, the largest measured decline in test scores across all districts, 12.4 percent for test score ratio in reading, occurred in District 5. District 5 has an enrollment share of 18.7 percent, the second largest in the state. Thus, negative effects in this district drag down state averages. District 4, the second largest minority district with 44.4 percent nonwhite enrollment, shows a similar pattern, with test score gains in language being offset by test score losses in reading and math. Districts 7 and 8 with nonwhite enrollments of 14.1 percent and 15.8 percent, respectively, are the only districts in the state to show clear gains in test scores during the NCLB era. But here again, the gains are limited to language, and in the case of District 7, test score ratio for math.

5. Concluding comments

Study results based on a dynamic panel data model suggest Alabama's county schools adjust relatively quickly to exogenous shocks, within three years or less. Moreover, most of the adjustment, 75 percent or more, occurs in the first year of the policy innovation. This suggests interventions to improve academic performance can produce results in a relatively short period of time. As for the No Child Left Behind Act of 2001, model results suggest revisions in the state's accountability system associated with the Act had positive effects on SAT test scores for language. However, the measured effects on test scores for reading and math were mostly zero or negative. This suggests that in the case of Alabama's county schools, which are largely rural, NCLB has failed in its main aim of improving proficiency in reading and math, at least at the 8th grade level, and at least as measured through the 2007-08 school year. One reason for the failure may be that Alabama had an accountability system in place prior to NCLB. This system, which was

implemented in the 1995-96 school year, was given high marks by Carnoy and Loeb (2002) in their assessment of accountability systems nationwide. As noted by Lee (2006), in instances where states already have in place a “consequential” accountability system, NCLB is not apt to have much effect. This inference notwithstanding, our results did show a significant positive effect for language.

A criticism of NCLB is that the Act imposed demanding standards on the states without providing the necessary resources for their effective implementation (Lee, 2006, p. 5). In the case of Alabama’s county schools, this criticism appears not to apply, as per pupil spending between the pre- and post-NCLB periods (herein defined to be 1999-2002 and 2003-2007, respectively) rose 13.6 percent in real terms. Unfortunately, the increased spending was devoted to inputs other than teacher pay, and thus, according to our econometric results, had no effect on test scores. In fact, teacher pay, a key variable affecting test scores in our model, declined by 11.4 percent in real terms between the pre- and post-NCLB periods. Had this decline not occurred, model simulations suggest the observed decline in 8th grade test scores between the periods would have been 29 percent less for reading, and 25 percent less for math. In revising its state accountability system to bring it into compliance with NCLB, Alabama developed a new test, the Alabama Reading and Math Test, for the purposes of meeting the Adequate Yearly Progress provision of the federal law. Although the ARMT was implemented in 2004, the state continued to administer the Stanford Achievement Test, which this study relies on for its performance data. The shift in emphasis in the state from the SAT to the ARMT as the basis for accountability reduces to some extent concerns about “teaching to the test.”⁹ Still, even in the best of circumstances scores on standardized tests are imperfect indicators of academic achievement, a caveat that must be borne in mind when interpreting the results of this study.

References

- Alabama Department of Education. (2007a). *Annual Report, 2006-2007*. Montgomery, AL.
- Alabama Department of Education (2007b). *Schools in School Improvement:Status for 2007-2008*.
- Alabama Department of Education (2007c). *Schools that Did Not Make AYP Status for 2007-2008*.
- Alabama Department of Education.(2010). *2010 Interpretive Guide:Alabama Accountability System*.
- Arellano,M., Bond, S.R. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277-97.
- Bond, S. R. (2002). Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal*, 1, 141-62.
- Code and Alabama. (2001 replacement). *Education Accountability Plan*. Title 16, Chapter 6B.
- Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics*, 30, 109-26.
- Dee, T. and Jacob, B. (2009). *The impact of No Child Left Behind on student achievement*. Working Paper 15531. National Bureau of Economic Research. Cambridge, MA.
- Dee, T. S. and Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30, 418-446.
- Duffy, M. (1989). Measuring the contribution of advertising to growth in demand: an econometric accounting framework. *International Journal of Advertising*, 8, 95-110.

⁹ According to press releases from the Alabama Department of Education, available on its web site, the share of public schools making AYP between 2003 and 2007 increased from 31 percent to 82 percent. This suggests scores on the ARMT over this period were increasing, an outcome that is not fully consistent with the SAT data used in this study. The point being that “teaching to the test” is a legitimate concern.

- EViews 7. (2009). Quantitative micro software, LLC. Irvine, California.
- Ferguson, R. F. and Ladd, H. F. (1996). Additional evidence on how and why money matters: a production function analysis of Alabama schools." Chapter 8 (pp. 265-297) in *Holding Schools Accountable: Performance-Based Reform in Education*. H. F. Ladd (editor). Washington, D.C.: Brookings Institution.
- Figlio, D. N., & Ladd, H. (2008). *School accountability and student achievement*. In H. Ladd & E. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 166–182). New York: Routledge.
- Griliches, Z. (1967). Distributed lags: a survey. *Econometrica*, 35, 16-49.
- Halvorsen, R. and Palmquist, R. (1980). The interpretation of dummy variables in semilogarithmic equations. *American Economic Review*, 70, 474-75.
- Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, 57, 670-94.
- Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: an update. *Educational Evaluation and Policy Analysis*, 19, 141-164.
- Hanushek, E. A. (1986). The economics of schooling: production and efficiency in public schools. *The Journal of Economic Literature*, 24, 1141-1177.
- Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management*, 1, 19-41.
- Hanushek, E. A. (2004). What if there are no 'best practices'? *Scottish Journal of Political Economy*, 51, 157-172.
- Hanushek, E. A. and Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24, 297-327.
- Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100, 267-71.
- King, R. G. and Thomas, J. K. (2006). Partial adjustment without apology. *International Economic Review*, 47, 779-809
- Kinnucan, H. W, Zheng, Y. and Bremer, G. (2006). State aid and student performance: a supply-demand analysis. *Education Economics*, 14, 487-509.
- Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Boston: Civil Rights Project, Harvard University.
- Levačić, R. and Vignoles, A. (2002). Researching the links between school resources and student outcomes in the UK: A review of issues and evidence. *Education Economics*, 10, 313-331.
- Meyer, R. H. (1997). Value-added indicators of school performance: a primer. *Economics of Education Review*, 16, 283-301.
- Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92, 263–283.
- Nerlove, M. and Bessler, D. A. (2001). Expectations, information and dynamics. In *Handbook of Agricultural Economics*. Volume 1, Part 1. Pp. 155-206.
- Pritchett, L. and Filmer, D. (1999). What education production functions *really* show: a positive theory of education expenditures. *Economics of Education Review*, 18, 223-39.
- Public Affairs Research Council of Alabama. (1997). Looking at Alabama's public school investment and performance. parca.samford.edu/PARCA2/execsummaries/educate.pdf (accessed 15 August 2011).
- Rothstein, J. (2010). Teacher quality in education production: tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125, 175-214.
- Smith, D. S. (2009). *An Analysis of the No Child Left Behind Act Using Gradual Switching Regressions*. M. S. Thesis. Auburn University, Auburn, Alabama.

Appendix 1. Tables

Table 1. Summary statistics for variables used in the model (upper number is for 1999-2002 (*nobs* = 268); lower number is for 2003-2007 (*nobs* = 335))

Variable	Definition	Mean	SD	Min	Max
RTS8	Reading 8 th grade test score, national percentile ^a	47.2 43.6	9.1 9.0	20 21	68 66
MTS8	Math 8 th grade test score, national percentile ^a	47.1 44.8	9.3 9.3	23 21	70 70
LTS8	Language 8 th grade test score, national percentile ^a	48.3 47.4	6.8 7.7	28 27	65 69
RTS4	Reading 4 th grade test score, national percentile ^a	49.1 52.1	10.4 10.4	17 24	70 72
MTS4	Math 4 th grade test score, national percentile ^a	53.5 54.1	8.7 9.0	26 32	67 73
LTS4	Language 4 th grade test score, national percentile ^a	61.9 60.4	6.8 8.2	36 38	77 81
PPS	Per pupil spending, 1982-84 dollars	3,801 4,318	426 845	3,053 3,185	5,891 10,222
TPAY	Annual teacher pay, 1982-84 dollars	22,463 19,899	818 2,479	20,262 13,199	26,090 25,781
ENROL	School district size, average daily membership	7,248 7,287	9,851 9,902	1,573 1,124	65,235 65,686
RACE	Nonwhite student enrollment, percent	38.4 39.1	31.0 30.9	0.1 0.3	100 100
INC	Median family income, 1982-84 dollars	17,556 17,179	3,726 3,682	10,259 9,895	34,524 33,095
POV	Poverty rate, percent	22.6 24.8	5.9 7.2	7.3 7.4	41.8 43.3
RURAL	Rural county, dummy variable	0.74 0.72	0.44 0.45	0 0	1 1
UNEMP	County unemployment rate, percent	5.8 5.0	2.0 1.7	1.5 2.3	13.0 11.3
PV	Property value per pupil, 1982-84 dollars	31,725 37,679	16,668 19,414	10,831 13,213	87,265 107,616
ED	Adults 25 and older with at least a Bachelor's degree, percent	13.6 14.6	6.3 6.8	6.9 6.1	38.4 42.3
	Dummy variables:				
NCLB	No Child Left Behind	0.56	0.50	0	1
D1	State Board of Education District 1	0.045	0.21	0	1
D2	State Board of Education District 2,	0.194	0.40	0	1
D3	State Board of Education District 3	0.134	0.34	0	1
D4	State Board of Education District 4	0.045	0.21	0	1
D6	State Board of Education District 6	0.060	0.24	0	1
D7	State Board of Education District 7	0.149	0.36	0	1
D8	State Board of Education District 8	0.075	0.26	0	1

^a Scores are based on the Stanford Achievement Test, 10 edition.NOTE: Sources provided in

Appendix A.

Table A1B. Mean scores on Stanford Achievement Test before and after No Child Left Behind Act, Alabama county schools

Subject	4 th Grade				8 th Grade		
	Before NCLB (1999-2002)	After NCLBE (2003-2007)	Percent Change		Before NCLB (1999-2002)	After NCLBE (2003-2007)	Percent Change
Reading	49.2	52.1	5.9		47.2	43.6	-7.6
Math	53.5	54.1	1.1		47.1	44.8	-4.9
Language	61.9	60.4	-2.4		48.3	47.4	-1.9
All	54.9	55.5	1.1		47.6	45.3	-4.8

Table A2

Variable/ Statistic	Reading		Math		Language	
	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
C	-5.521	-3.10	-0.797	-0.45	-6.870	-4.50
LOG(TS8th(-1))	-0.027	-0.63	0.197	4.88	0.102	2.16
NCLB	-0.064	-4.10	-0.034	-2.51	0.009	0.52
NCLB x D1	0.056	2.79	0.040	1.94	0.044	1.50
NCLB x D2	0.035	1.64	0.031	1.51	0.011	0.44
NCLB x D3	0.040	1.53	0.030	1.26	0.040	1.66
NCLB x D4	-0.004	-0.22	-0.020	-1.21	0.007	0.26
NCLB x D6	-0.006	-0.15	0.009	0.36	0.014	0.44
NCLB x D7	0.048	2.13	0.034	1.95	0.055	2.67
NCLB x D8	0.046	1.80	0.019	0.76	0.030	1.10
LOG(INC)	0.488	5.20	0.166	1.73	0.259	3.09
LOG(POV)	-0.006	-0.14	-0.088	-2.21	-0.122	-2.23
RURAL	-0.040	-1.56	-0.050	-1.84	-0.026	-1.39
LOG(RACE)	0.004	0.12	-0.110	-2.78	-0.027	-0.57
LOG(DSIZE)	0.404	3.72	0.243	1.90	0.498	4.32
LOG(PPS)	-0.079	-0.88	0.005	0.07	0.109	1.12
LOG(TPAY)	0.197	3.88	0.082	1.66	0.318	5.40
R ² /Adjusted R ²	0.919	0.904	0.910	0.891	0.856	0.830
S.E. of regression/RSS	0.07321	2.4279	0.08772	3.4851	0.07315	2.4238
D.W. statistic/J-statistic	2.09	3.25	2.07	4.64	2.08	2.42
Instrument rank/NOBS	86	536	86	536	86	536

Table A3

Variable/ Statistic	Reading		Math		Language	
	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
C	-7.713	-3.83	-7.072	-2.55	-8.621	-5.77
LOG(TS8th(-1)/TS4th(-1))	0.084	1.68	0.236	5.59	0.109	2.31
NCLB	-0.114	-7.03	-0.053	-2.24	0.032	2.56
NCLB x D1	0.078	3.52	0.014	0.24	0.068	1.97
NCLB x D2	0.046	1.91	0.092	2.86	0.048	2.21
NCLB x D3	0.071	2.53	0.052	1.73	0.035	1.81
NCLB x D4	0.076	4.40	0.056	2.16	0.077	1.96
NCLB x D6	0.043	1.40	0.036	0.86	0.020	0.52
NCLB x D7	0.081	3.03	0.114	3.74	0.060	2.81
NCLB x D8	0.090	2.88	0.068	1.56	0.051	1.87

LOG(INC)	0.224	1.78	-0.318	-1.87	0.135	1.40
LOG(POV)	-0.129	-3.58	-0.275	-2.85	-0.159	-2.26
RURAL	-0.106	-3.31	-0.099	-2.08	-0.084	-2.65
LOG(RACE)	0.070	2.04	-0.026	-0.51	-0.014	-0.30
LOG(DSIZE)	0.241	1.84	0.320	2.83	0.247	3.68
LOG(PPS)	-0.010	-0.12	0.255	1.49	0.086	0.71
LOG(TPAY)	0.378	6.05	0.626	5.26	0.488	6.98
R ² /Adjusted R ²	0.700	0.641	0.711	0.656	0.672	0.613
S.E. of regression/RSS	0.0983	4.377	0.1229	6.841	0.0908	3.740
D.W. statistic/J-statistic	2.14	3.04	2.04	3.86	2.13	3.43
Instrument rank/NOBS	86	536	86	536	86	536

Table 4. Estimated effects of changes in model variables on 8th grade test scores

Variable	Long-Run Elasticity			Level of Variable		Percent Change in Variable	Predicted Change in Test Score (Percent)		
	Reading	Math	Language	1999-2002	2003-2007		Reading	Math	Language
Family income	0.488	0.206	0.288	17,556	17,179	-2.15	-1.05	-0.44	-0.62
Poverty rate	0.000	-0.110	-0.135	22.6	24.8	9.73	0.00	-1.07	-1.31
Rural	-0.040	-0.062	-0.029	0.74	0.72	-2.70	0.11	0.14	0.08
Race	0.000	-0.137	0.000	38.4	39.1	1.82	0.00	-0.25	0.00
District size	0.404	0.302	0.554	7,248	7,287	0.54	0.22	0.16	0.30
Per pupil spending	0.000	0.000	0.000	3,801	4,318	13.60	0.00	0.00	0.00
Teacher pay	0.197	0.102	0.354	22,463	19,899	-11.41	<u>-2.25</u>	<u>-1.16</u>	<u>-4.04</u>
						Sum:	-2.97	-2.62	-7.62
						Observed Change	-7.62	-4.88	-1.86

Table 5. Estimated effects of changes in model variables on 8th/4th grade test score ratios

Variable	Long-Run Elasticity			Level of Variable		Percent Change In Variable	Predicted Change in Test Score Ratio (Percent)		
	Reading	Math	Language	1999-2002	2003-2007		Reading	Math	Language
Family income	0.244	-0.416	0.000	17,556	17,179	-2.15	-0.52	0.89	0.00
Poverty rate	-0.141	-0.360	-0.178	22.6	24.8	9.73	-1.37	-3.50	-1.73
Rural	-0.116	-0.130	-0.095	0.74	0.72	-2.70	0.31	0.35	0.26
Race	0.076	0.000	0.000	38.4	39.1	1.82	0.14	0.00	0.00
District size	0.263	0.418	0.277	7,248	7,287	0.54	0.14	0.23	0.15
Per pupil spending	0.000	0.000	0.000	3,801	4,318	13.60	0.00	0.00	0.00

Teacher pay	0.413	0.820	0.548	22,463	19,899	-11.41	<u>-4.71</u>	<u>-9.36</u>	<u>-6.25</u>
						Sum:	-6.02	-11.39	-7.58
						Observed Change	-7.62	-4.88	0.57

The inimical effects of reduced teacher pay are even more pronounced for test score ratio. Here, the decline in test score gains for reading, math, and language associated with the 11.4 percent reduction in teacher pay are, respectively, 4.7 percent, 9.4 percent, and 6.2 percent. The observed changes in test score gains for reading, math, and language are, respectively, -7.6 percent, -4.9 percent, and 0.6 percent. This suggests, for example, that if teacher pay between the periods had merely stayed constant in real terms, test score gains in math would have increased rather than decreased. This observation is poignant because per pupil spending, which has no effect on test scores, increased between the periods by 13.6 percent in real terms. Hence, monies were available to augment teacher pay, they just were not used for that purpose, a classic allocation error.

Results indicate that both for test score level (table 4) and test score ratio (table 5), the most influential variable by far is *TPAY*. Real teacher pay between the periods declined 11.4 percent. This reduction alone is associated with decreases in 8th grade test scores for reading, math, and language of 2.2 percent, 1.2 percent, and 4.0 percent, respectively. By way of comparison, the observed reductions in test scores for reading, math, and language are 7.6 percent, 4.9 percent, and 1.9 percent. This suggests 29 percent of the observed decline in reading scores and 25 percent of the observed decline in math scores is attributable to the decline in teacher pay.

At the same time average real teacher pay decreased by 11.4 percent, its standard deviation increased by 303 percent (from \$818 to \$2,479, see table 1). This suggests Alabama's county school systems are willing to tolerate larger differences in teacher pay, which opens up the possibility of rewarding teachers based on performance. If this is true, the county school systems would do well to devote incremental funds to increases in teacher pay that are based on merit rather than seniority or other criteria unrelated to performance.

Appendix 2. **Alabama accountability**

In 1995, the Alabama Legislature passed the *Education Accountability Plan* (Code of Alabama, 2001 Replacement). The plan, which went into effect in the 1995-96 school year, required all schools in the state be accountable in four areas: fiscal responsibility, school safety, discipline, and student achievement. Academic accountability worked as follows (PARCA, 1997, p. 3):

The student achievement standards require that all students in grades 3-11 take the norm-referenced Stanford Achievement Test (SAT) each year for accountability purposes. SAT scores are presented in terms of a percentile rank that compares an Alabama score with a national sample of students who took the test in 1995. The percentile ranks range from 0 to 100, and the national average is 50. Each school and school system is graded on the basis of its students' scores.

Any school or system with a majority of students scoring below the 23rd percentile goes on academic alert and must reduce the number of students scoring below this level by 5

percentage points each year. If this progress is not made, the school or system faces state intervention after 3 years. Schools on alert conduct self assessments and receive state assistance. Each year the State Department of Education hires a number of outstanding Alabama teachers on one-year contracts to work with schools that need help. Any school or system not on alert, but with a majority of students scoring below the 40th percentile, goes on academic caution and must reduce the number of students scoring below this level by 2 percentage points. If this progress is not made in the next year, the school or system drops into alert status and must meet the criteria for that category as discussed above. Schools with a majority of students scoring at or above the 40th percentile are rated clear.

Core elements of NCLB were in place in Alabama prior to the federal law. The Alabama system was deemed “strong” by Carnoy and Loeb (2002) in their evaluation of accountability systems nationwide.¹⁰ Still, the federal law posed significant new challenges, and it took until June 2003 for the State Board of Education to adopt a single accountability system that was in compliance with NCLB and state law (ADE 2010). The centerpiece of the new accountability system was procedures for meeting requirements for AYP. A criterion-referenced test called the Alabama Reading and Mathematics Test (ARMT) was developed to assess students’ progress in meeting state-based standards in reading and mathematics (ADE 2010). The ARMT, which supplements the norm-referenced SAT test, was administered for the first time in spring 2004. Since then the test has been given every year to public-school students in grades 3 through 8, and grade 11.

Results of the ARMT are reported as percentages of students that exceed (Level IV), meet (Level III), partially meet (Level II), and do not meet (Level I) academic content standards. Students scoring at Level III or IV are considered proficient. The goal of NCLB is for all public school students to perform at the proficient level by the end of the 2013-14 school year. Baseline, or starting points, for meeting this goal were established for each school based on test results from the ARMT administered in the spring of 2004 and 2005.

AYP is determined based on three criteria: participation rate, annual measurable objectives in reading and mathematics, and additional academic indicators. To satisfy the participation rate (PR) criterion, the ARMT must be taken by at least 95 percent of enrolled students. To satisfy the annual measurable objectives (AMO) criterion, the percentage of students scoring Level III or IV on the ARMT must increase each year in accordance with intermediate goals established on the basis of baseline scores. To satisfy the additional academic indicator (AAI) criterion, schools must show an attendance rate of at least 90 percent. If the school has grade 12 (meaning it confers a high school diploma), it must also show a 90 percent graduation rate. For a school to make AYP, all three criteria (PR, AMO, and AAI) must be satisfied.

Schools that do not make AYP are posted on the Alabama Department of Education’s website. If a school does not make AYP for two consecutive years, it is identified for

¹⁰ On Carnoy and Loeb’s zero-to-five “strength of accountability index” Alabama received a “4,” one of only nine states to score as high. The other states scoring 4 or above are California, Kentucky, Maryland, New Jersey, New Mexico, New York, North Carolina, and Texas (Carnoy and Loeb, 2002, pp. 324-326).

“School Improvement Year 1.” If the school does not make AYP in the next year, it moves to “School Improvement Year 2,” and so forth. Once in School Improvement, if a school makes one year AYP, it does not advance to the next level, but retains its current improvement status, also known as the “delay provision.” In 2007, the last year of our sample, 16.3 percent of Alabama’s public schools did not make AYP, and 10.8 percent were in School Improvement (ADE 2007bc).

A major concern in using test scores to evaluate student performance is that they may reflect “teaching to the test” rather than actual gains in cognitive ability or academic achievement (see, e.g., Neal and Schanzenbach 2010 and references therein). Indeed, after the 1995 state accountability system was implemented, average scores on the Stanford Achievement Test statewide rose steadily so that by 1997 “every grade tested scored above the national average of 50” (PARCA, 1997, p. 4). Yet an examination of 8th grade test scores in math on the National Assessment of Educational Progress (NAEP) showed that in 1996 “Alabama students scored lower than their counterparts in 37 other states and finished ahead of only Louisiana and Mississippi” (PARCA, 1997, p. 5). Dee and Jacob (2011, p. 419) argue that “low-stakes” test data from the NAEP are more reliable than “high-stakes” data from state assessments. In the present study, we use SAT data to be consistent with the earlier studies by FL and KZB. The fact that the ARMT rather than the SAT is used to determine AYP mitigates to some extent concerns about teaching to the test.

In 2003, Alabama switched from the Stanford 9 to the Stanford 10 test. The Stanford 9 is based on 1995 norms and the Stanford 10, a different test, is based on 2002 norms. To make test scores comparable before and after 2003, we converted all Stanford 9 test scores to Stanford 10 normed scores using conversion tables developed Harcourt Assessment, the provider of the SAT test.¹¹

Appendix 3. Elasticities

Table 3. Estimated long-run elasticities from education production functions, Alabama county schools, 2000-2007 data

Education Input	8 th Grade Test Score Level			8 th /4 th Grade Test Score Ratio		
	Reading	Math	Language	Reading	Math	Language
<u>Accountability system:</u>						
<i>NCLB x D1</i>	-0.009 (-0.61) ^a	0.007 (0.35)	0.059** (2.11)	-0.040** (-1.98)	-0.051 (-0.66)	0.111** (2.87)
<i>NCLB x D2</i>	-0.029* (-1.93)	-0.003 (-0.16)	0.022 (1.02)	-0.073 (-3.94)	0.052* (1.80)	0.089** (4.28)
<i>NCLB x D3</i>	-0.024 (-1.23)	-0.005 (-0.24)	0.054** (3.51)	-0.047* (-1.88)	-0.002 (-0.05)	0.074** (3.83)
<i>NCLB x D4</i>	-0.068** (-5.78)	-0.066** (-4.63)	0.017 (0.87)	-0.041** (-2.50)	0.004 (0.14)	0.123** (2.90)
<i>NCLB x D5</i>	-0.064** (-4.10)	-0.042 (-1.29)	0.010 (0.51)	-0.124** (-6.40)	-0.069** (-2.13)	0.036** (2.62)

¹¹ All data used in this study, including web-linked sources, are provided in an appendix available upon request.

<i>NCLB x D6</i>	-0.070** (-1.97)	-0.031 (-0.93)	0.026 (0.95)		-0.077** (-2.87)	-0.022 (-0.64)	0.058 (1.51)
<i>NCLB x D7</i>	-0.016 (-0.90)	0.0008 (0.05)	0.071** (4.67)		-0.036 (-1.49)	0.079** (2.75)	0.103** (4.70)
<i>NCLB x D8</i>	-0.018 (-0.96)	-0.018 (-0.71)	0.044** (2.05)		-0.026 (-0.90)	0.019 (0.40)	0.093** (3.20)
Family and nonschool inputs (X variables):							
<i>INC</i>	0.488** (5.20)	0.206* (1.74)	0.288** (3.04)		0.244* (1.82)	-0.416* (-1.84)	0.151 (1.39)
<i>POV</i>	-0.006 (-0.14)	-0.110** (-2.22)	-0.135** (-2.17)		-0.141** (-3.53)	-0.360** (-2.81)	-0.178** (-2.21)
<i>RURAL</i>	-0.040* (-1.56)	-0.062* (-1.84)	-0.029* (-1.38)		-0.116** (-3.36)	-0.130** (-2.11)	-0.095** (-2.69)
<i>RACE</i>	0.004 (0.12)	-0.137** (-2.74)	-0.030 (-0.56)		0.076** (2.06)	-0.034 (-0.51)	-0.016 (-0.30)
School inputs (R variables):							
<i>DSIZE</i>	0.404* (3.72)	0.302* (1.92)	0.554** (4.31)		0.263* (1.81)	0.418** (2.92)	0.277** (3.72)
<i>PPS</i>	-0.079 (-0.88)	0.007 (0.07)	0.122 (1.09)		-0.011 (-0.11)	0.334 (1.49)	0.097 (0.70)
<i>TPAY</i>	0.197** (3.88)	0.102* (1.66)	0.354** (4.76)		0.413** (5.95)	0.820** (5.54)	0.548** (6.34)

^aNumber in parenthesis is asymptotic t-ratio. The delta method (with analytic derivatives) is used to compute the Wald restriction variance for the non-linear restriction associated with the lagged dependent variable. Double asterisk (**) indicates estimated coefficient is significant at the $p \leq 0.05$ level; single asterisk (*) indicates significance at $p \leq 0.10$ level. For full regression results, see appendix tables 2 and 3.

Of the 48 estimated policy responses, 24 are zero (at the $p < 0.10$ level), 11 are negative, and 13 are positive. Of the 13 positive responses, 11 are for Language. Of the 11 positive responses for language, seven are for test-score ratio. Thus, to the extent NCLB can be associated with positive effects, the effects are limited to language, and appear to be more important for test-score gain than for test-score level. Given that the goal of NCLB was enhance proficiency in reading and math, there is little in the SAT data to suggest this has occurred in Alabama's county schools. Math does show a positive response for test score ratio in District 2 (0.052) and District 7 (0.079). But these positive responses are offset by a negative response in District 5 (-0.069).

For all 48 responses, the largest estimated response is for test-score ratio for language in District 4 (0.123). In this instance, NCLB is associated with a gain in language SAT test scores of 12.3 percent. Our overall conclusion based on these estimates is that while NCLB may have had a positive effect on language proficiency (as measured by SAT scores), its effect on reading and math proficiency is, for the most part, nil or negative.

Turning to the X and R variables, and focusing first on test score level, most of the estimated elasticities have expected signs and are statistically significant (table 3). According to these estimates, the three most important variables affecting test score level are *DSIZE* (0.30 to 0.55), *INC* (0.21 to 0.49) and *TPAY* (0.10 to 0.35). The estimated elasticities for *POV*, *RACE* and *RURAL* are mostly statistically significant, but are too

small (less than 0.14 in absolute value) to be empirically significant. *PPS* is not statistically significant in any of the equations. One interpretation of this result, suggested by Pritchett and Filmer (1999), is that per pupil spending is optimized in the sense that its marginal product is zero. Another interpretation, consistent with Hanushek (1981), is that simply “throwing money at schools” is not apt to be effective in improving student performance. By way of comparison, KZB found the three most important variables affecting test scores are *POV* (-0.36), *SF* (0.23), and *INC* (0.20).¹² Hence, while family income remains an important determinant of student performance in Alabama’s county schools, it appears that poverty is no longer the impediment to student performance it once was. Recalling KZB’s injunction that increases in state aid is effective only if it is directed at teacher pay, it appears that this input remains important. The entry of district size into the list of important determinants suggests district size is not to be overlooked as one avenue for possible improvements in student performance. In terms of NCLB, it appears that county schools in larger districts have an advantage, *ceteris paribus*, in making AYP. Overall, test score responses are highly inelastic. Thus, while county schools respond to exogenous shocks fairly quickly, the responses themselves are modest.

Turning to test score ratio, the three most important variables are *TPAY* (0.41 to 0.82), *DSIZE* (0.26 to 0.42), and *POV* (-0.14 to -0.36). Thus, while poverty is not very important in explaining test score *level*, it appears to be relatively important in explaining in test score *gain*. Stated differently, poverty appears to be a relatively important factor in explaining the decline in student performance as students move through the county school system. Because *TPAY* and *DSIZE* are important both for test-score level and test-score gain, these variables should serve as the focal point for policy purposes. The elasticities for *RURAL* (-0.10 to -0.13) are negative and significant for all three subjects. This suggests students in rural counties suffer a 10 to 13 percent performance gap relative to their peers in non-rural counties. The elasticities for *INC* and *RACE* are mostly insignificant at the $p < 0.05$ level. The one exception is *RACE* for reading, where the estimated elasticity is 0.076. This suggests an isolated 10 percent increase in non-white student enrollment is associated with a 0.76 percent increase in the 8th/4th grade test score ratio for reading, *ceteris paribus*. Although the effect is tiny, it is interesting that it is positive rather than negative. The weak effects of the *RACE* variable across all six equations suggest race-based discrimination in Alabama’s county schools is a thing of the past. As with test score level, *PPS* is not significant in any of the test score ratio equations. Combining this result with the corresponding result for test score level, it would appear that the observed decline in test scores between the pre- and post-NCLB periods must be due to factors other than a decline in school resources.

¹² Midpoint estimates are taken from table 5, row 7 of KZB (2006, p. 502). The *SF* variable corresponds to state funding, which in the present analysis is included in the *PPS* variable.

Appendix 4. Preliminary analysis.

In preliminary analysis, we used the estimator developed by Arellano and Bond (1991) for situations where the panel has a large cross-section dimension and a short time series. (After accounting for the loss of one year due to the LDV, our panel has 67 counties and 8 years.) As explained by Bond (2002), when the time series is short, initial conditions play a larger role in the properties of estimators. Failure to account for this fact can cause the estimated coefficient of the LDV to be severely biased. Investigating this issue, we found little difference between the AB estimator and the GMM fixed-effects estimator. Because the GMM fixed-effects estimator produced more plausible estimates for the remaining parameters in the models, we elected to use that estimator. The fixed-effects estimator adds 67 instruments to the original instrument list (one for each cross section), resulting in a total of 86 instruments. Hence, the estimated models are over identified. The validity of the over-identifying restrictions is tested using the Sargan statistic.

The test scores in our panel are averages computed across schools in each county. As noted by Deaton (1985, p. 116), this averaging process can induce heteroscedasticity when cells sizes are very different. In our data, the number of schools per county ranges from 10 to 117, with a mean of 15. Given the large variation in cell size, we used the “cross-section weights” option in the EVIEWS to obtain standard errors that are robust to heteroscedasticity. Also, since policy shocks can induce within-group serial correlation (Hansen 2007), we used the “White period correction” option in EVIEWS (with no correction for degrees of freedom) to obtain standard errors that are robust to serial correlation.

Journal published by the EAAEDS: <http://www.usc.es/economet/eaat.htm>